Data analysis assignment 1 (1 week)

Imagine a startup company that has a few ideas for good apps and wants to make some moves on the app market. Before they can decide which of their ideas to start with, they need to have a better understanding of the market as it is now. They have used an API to collect some generally available data from the app store, but have no idea how to properly evaluate it. That's where you come in. Use the provided dataset to create valuable insights for the company to help them in deciding the best direction to take. If you're having troubles with anything, you can ask for hints on how to tackle the problem.

- Start with using a Python IDE to load in the 'googleplaystore.csv' dataset. For this I recommend Jupyter Notebooks (which can be acquired by downloading the Anaconda package) or Google Collab.

- Next, think about possible ways to clean up the dataset. The dataset is relatively clean, but there might be some columns that need to be transformed a bit to make them easier to work with. In an actual company environment no one will tell you when exactly a dataset is "clean", so this is always up for interpretation. For example null-values: do you remove the rows containing them or do you fill them with a value that makes sense (either by interpolation or by filling them with a constant) or maybe just leave them in? Some other things you can do to clean up the dataset is to convert the numeric columns to the correct datatype. For example the column size contains values like '3M'. This is a string and hard to work with, while in essence it's a numeric value. To be a able to convert it to a numeric (int/float), you need to remove the 'M'. But now the value is 3, so how do you handle a size of '50K'? Same for the "Price" column. This column also makes sense to have as a numeric, but this column contains dollar signs. To turn this into a numeric column, the dollar signs need to be removed.

- Create some Explanatory Data Analysis as you inspect the dataset. Maybe see if you can find some interesting relationships between some of the columns? You can do this i.e. by creating a correlation matrix or some pairplots.

- Next up you can repeat the previous steps, but now for the 'googleplaystore_user_reviews.csv' dataset. Try to match both datasets. On what column are you joining them? Is there missing information? Does this provide additional information?

- Finish by creating a notebook with a few visualizations that tells the company all they need to know about the market according to you. Think about this as if you were to design a dashboard for management. What to put in and what to leave out? This should probably contain some high level overviews like averages and total statistics. And then gradually go towards deeper and more detailed insights like correlations and interesting finds. Also give an explanation on why you choose to display each visual and what the interesting thing is. In a company environment choosing the correct visuals and information to show is an iterative process, so try to be creative and remember it doesn't have to be perfect on the first try.